**Project Report**

Group 8

Krishna Yashwanth Tummala

Pramit Mukherjee

857‐869‐2088 (Tel of Student 1)

857‐222‐2831 (Tel of Student 2)

tummala.k@husky.neu.edu

mukherjee.p@husky.neu.edu

**Percentage of Effort Contributed by Student 1:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Percentage of Effort Contributed by Student 2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature of Student 1:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature of Student 2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Submission Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Project Proposal**

**Problem Setting:** Mt. Rainier is among America’s most beautiful and tallest mountain. Many expeditions are set out to scale the summit but only few are successful. The aim of this project is to model a machine learning algorithm that will help us in predicting the success of a climb given the weather conditions on that day. A variety of factors most notably the weather and the route of the climb play important roles in dictating the success of a climb.

**Problem definition:** The aim of this project is to reduce uncertainty in making a climb. The model will be able to successfully predict/classify a climb as success or failure by modeling the relationships between the predictors and the output variables. Analyzing the predictions of the model along with proper domain knowledge in mountaineering would considerably negate the risk of dangerous accidents and make the activity much safer.

**Data Sources:** The data is gathered from several sources, mainly the dataset from Kaggle on Mount Rainier weather and climbing data. The weather has been captured from [https://www.nwac.us](https://www.nwac.us/) and the climbing statistics from <http://www.mountrainierclimbing.us/routes>. Information about the mountain has been obtained for Google and Wikipedia.

**Data Description:** The data from Kaggle includes weather observation recorded at Mt. Rainier for the calendar years, 2014 and 2015. For each day of the year, the average battery voltage in volts, Average temperature in F, Average relative humidity in %, Wind speed in MPH, Wind direction in degrees and Solar radiation in W/m2 are recorded. The dataset also contains climbing statistics for the mountain from 2014-15. The climbing statistics contain the routes taken, number of attempts and successes and the success ratio for each day of the 2-year period. The output variable can be the success percentage which is numeric, or it can be transformed into a categorical variable that indicates if a climb will be successful.

**Data Collection**

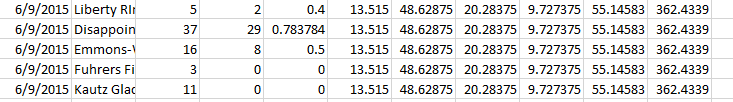
We have collected the weather data recorded at Mt.Rainier for the calendar years 2014 and 2015. The weather data includes Average battery voltage in Volts, Average temperature in degrees F, Average relative humidity in %, Average wind speed in MPH, Wind direction in degrees and Average solar radiation in W/m2. The climbing data includes climbing statistics for each day. This data includes information on different routes taken, number of attempts through the route, number if successes and the ratio of successes to attempts for that route on that day. Based on these observations we would be able to predict whether the climbers will be heading towards success or failure based on the route they take and the weather conditions. The predictors variables include the Average battery voltage in Volts, Average temperature in degrees F, Average relative humidity in %, Average wind speed in MPH, Wind direction in degrees and Average solar radiation in W/m2. We have collected the dataset from Kaggle.com and would like to thank the author of the dataset Mr. Sree Harsha for providing us with the same. The nature of our predictors is Quantitative, as our data deals mostly with numbers. Furthermore, a new classification output variable is introduced which take the values Yes for high success ratio and No for low success ratio. There will be a lot of missing values in the climbing statistics data because climbing is possible only one some days of the year based primarily on the weather conditions. We will be using both Nominal and Numerical output variables to test and improve on the performance of the model.

There are two types of quantitative data, which is also referred to as numeric data: ***continuous***and ***discrete****.*As a rule, *counts*are discrete, and *measurements*are continuous.

We plan to clean the data by imputing and missing values with zeros. This can be done because a missing value in the climbing statistics means climbs were not attempted which would equate to numeric zero when one climb takes the value numeric one. The climbing statistics include climbing observations for 4077 climbs attempted between 2014 and 2015 through various routes. Somedays will have success/attempt ratio more than one because some climbs are undertaken overnight while camping at a basecamp and completed the following day.

**Data Processing**

The weather data is missing for dates past 9/23/2014. To ensure the model does not encounter a problem, we limit the climbing statistics to the date above. The climbing statistics must be cleaned because multiple entries exist for the same date, sometimes multiple entries for the same route taken. Since we are using average temperatures, it would make sense to group the entries based on the date, that is if the climb entries fall on the same day for one of the routes taken. This would give us different entries for the same dates all grouped according to the routes used for attempting the climb. An example is shown in the figure below.



In the above figure, for the same date, 6/9/2015. 5 routes were used, and the attempts are grouped based on the route used, irrespective of the time they were started at. At the start of our analysis, we had 2360 rows of climb data based on the date and time of the climb. To incorporate use with average weather data, we need to group the climb based on the route. By grouping the data, we end up with 481 rows of data for climbs based on the data and the weather on that date. To make sure routes are differentiated and have an impact in the analysis, weather data for the same day is repeated for all routes. Now we have our dataset with 481 rows and 12 columns for the predictor and response variables, which can be used for modelling machine learning algorithms to predict/classify results.

**Data Exploration**

There are 4077 entries in the climbing statistics. Pivot Tables are extremely useful in viewing the information in the data in a meaningful way. By using pivot tables, we can view the climbing statistics for in a meaningful way. The climbing statistics for all dates are given below.

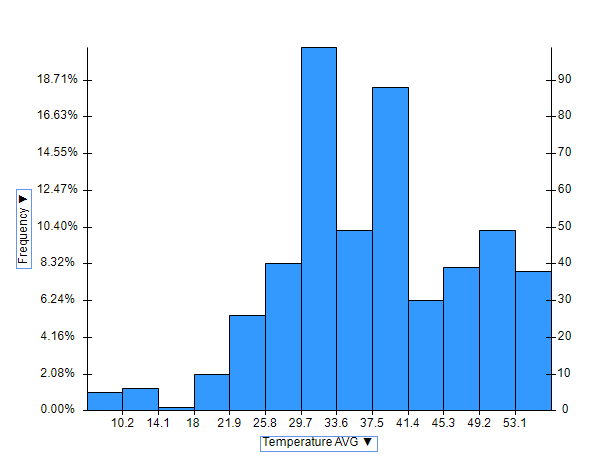
|  |  |  |  |
| --- | --- | --- | --- |
| **Row Labels** | **Sum of Attempted** | **Sum of Succeeded** | Ratio |
| Curtis RIngraham Directge | 4 | 2 | 0.5 |
| Disappointment Cleaver | 15259 | 8353 | 0.547414641 |
| Edmonds HW | 4 | 0 | 0 |
| Emmons-Winthrop | 3048 | 1673 | 0.548884514 |
| Fuhrers Finger | 76 | 9 | 0.118421053 |
| Fuhrer's Finger | 197 | 79 | 0.401015228 |
| Gibralter Chute | 13 | 2 | 0.153846154 |
| Gibralter Ledges | 199 | 58 | 0.291457286 |
| glacier only - no summit attempt | 187 | 15 | 0.080213904 |
| Ingraham Direct | 225 | 16 | 0.071111111 |
| Kautz Cleaver | 51 | 16 | 0.31372549 |
| Kautz Glacier | 949 | 530 | 0.558482613 |
| Kautz Headwall | 3 | 0 | 0 |
| Liberty RIngraham Directge | 165 | 82 | 0.496969697 |
| Liberty Wall | 2 | 0 | 0 |
| Little Tahoma | 296 | 135 | 0.456081081 |
| Mowich Face | 8 | 2 | 0.25 |
| Nisqually Glacier | 14 | 0 | 0 |
| Ptarmigan RIngraham Directge | 57 | 29 | 0.50877193 |
| Success Cleaver | 16 | 7 | 0.4375 |
| Sunset Amphitheater | 2 | 0 | 0 |
| Sunset RIngraham Directge | 6 | 0 | 0 |
| Tahoma Cleaver | 3 | 3 | 1 |
| Tahoma Glacier | 33 | 11 | 0.333333333 |
| Unknown | 131 | 30 | 0.229007634 |
| Wilson Headwall | 5 | 0 | 0 |
| **Grand Total** | **20953** | **11052** | 0.527466234 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Row Labels** | **Sum of Succeeded** | **Sum of Attempted** | **Average of Success Percentage** |
| **2014** | **6164** | **10926** | **0.530738693** |
| **Qtr1** | **33** | **155** | **0.226449275** |
| Jan | 16 | 64 | 0.289855072 |
| Feb | 0 | 28 | 0 |
| Mar | 17 | 63 | 0.220588235 |
| **Qtr2** | **1774** | **3795** | **0.445650615** |
| Apr | 6 | 157 | 0.043902439 |
| May | 477 | 1058 | 0.43072757 |
| Jun | 1291 | 2580 | 0.484087503 |
| **Qtr3** | **4356** | **6967** | **0.591726112** |
| Jul | 2288 | 3621 | 0.607584163 |
| Aug | 1542 | 2351 | 0.606766116 |
| Sep | 526 | 995 | 0.479721918 |
| **Qtr4** | **1** | **9** | **0.2** |
| **2015** | **4888** | **10027** | **0.461539015** |
| **Qtr1** | **19** | **192** | **0.109090909** |
| Jan | 10 | 42 | 0.230769231 |
| Feb | 3 | 56 | 0.05 |
| Mar | 6 | 94 | 0.090909091 |
| **Qtr2** | **2233** | **4212** | **0.48461045** |
| Apr | 19 | 116 | 0.122222222 |
| May | 413 | 1070 | 0.312294472 |
| Jun | 1801 | 3026 | 0.567266645 |
| **Qtr3** | **2636** | **5590** | **0.465959715** |
| Jul | 1626 | 2858 | 0.534998686 |
| Aug | 884 | 1959 | 0.446096399 |
| Sep | 126 | 773 | 0.166466132 |
| **Qtr4** | **0** | **33** | **0** |
| **Grand Total** | **11052** | **20953** | **0.498998836** |

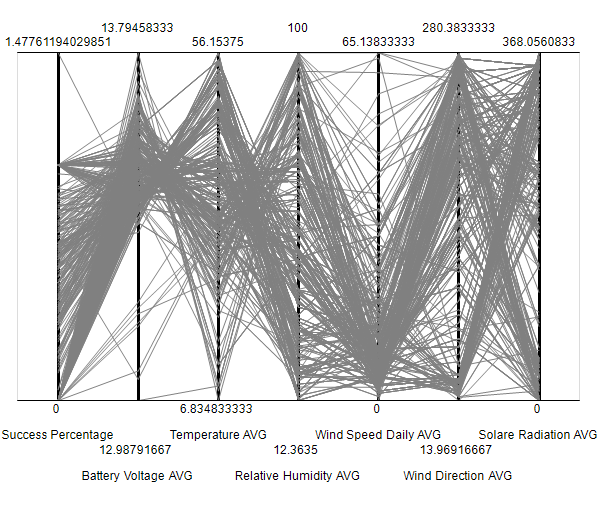
Similarly, the climbing statistics for 6/11/2015 are shown below

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Attempted** | **Sum of Succeeded** |
| Disappointment Cleaver | 51 | 18 |
| Emmons-Winthrop | 15 | 3 |
| Kautz Glacier | 6 | 0 |
| Little Tahoma | 2 | 2 |
| **Grand Total** | **74** | **23** |

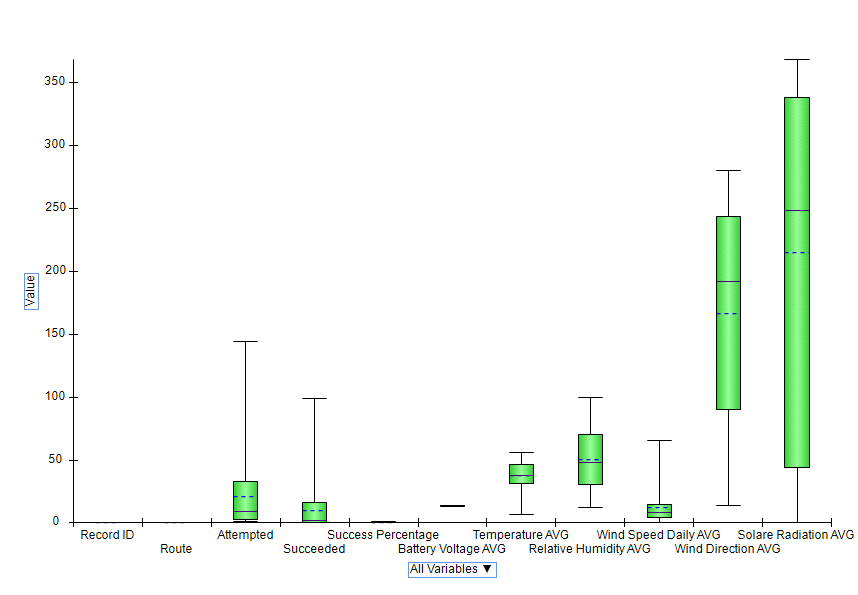
Histograms are used for studying the frequencies of the weather data. This will give us an idea for the more frequently encountered weather conditions. Two such histograms are shown below.



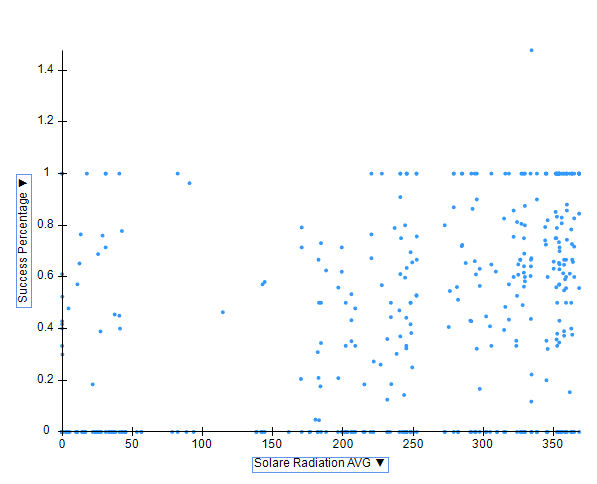
A parallel coordinates plot helps us in viewing the relationships among the predictor variables and the ratio of successes to attempts. This plot also gives us an idea of the extreme values for all the variables

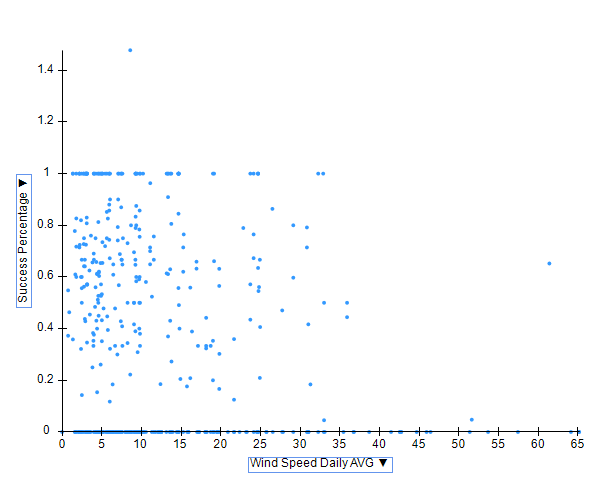


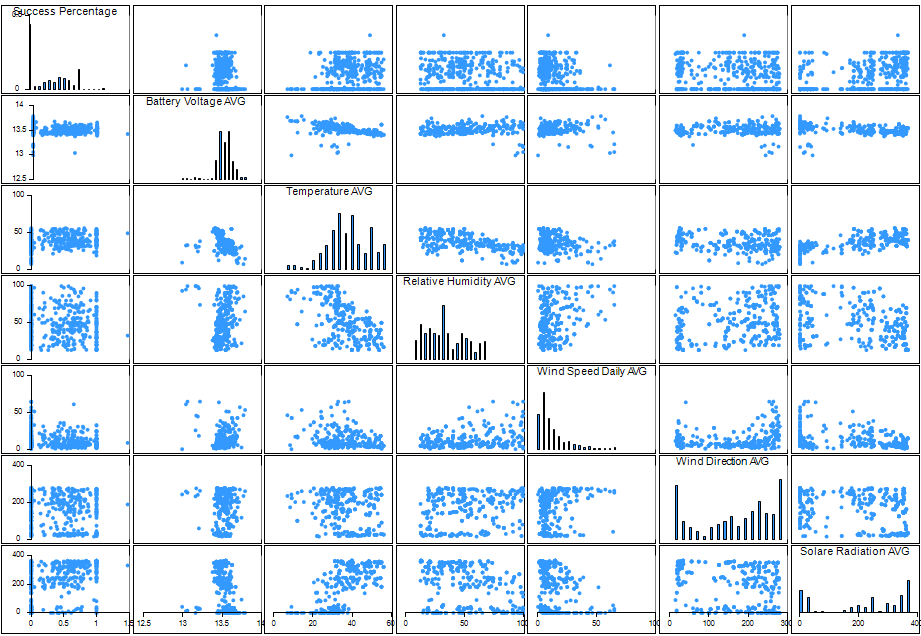
The boxplots help us in understanding the distribution of variables over the range of their values. We can see that the Solar radiation, Wind direction and relative humidity have comparatively higher distribution and thus will make excellent predictors.

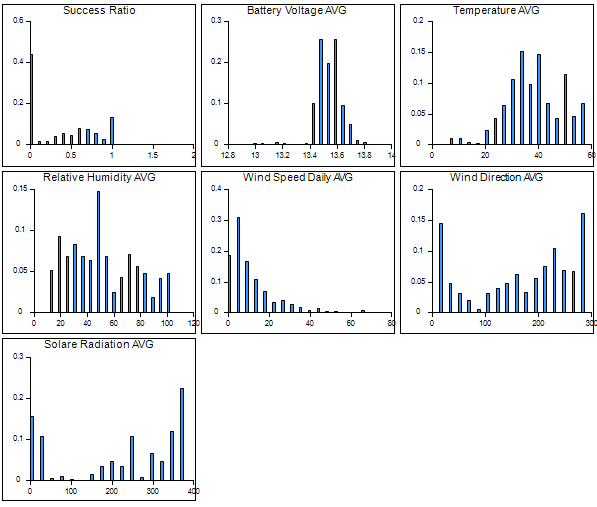


A scatterplot matrix helps us in viewing the relationships between the variables. Individual scatterplots help us in viewing the distribution on the success ratio with the individual predictor variables. A variable plot helps us in viewing all the variable distributions.







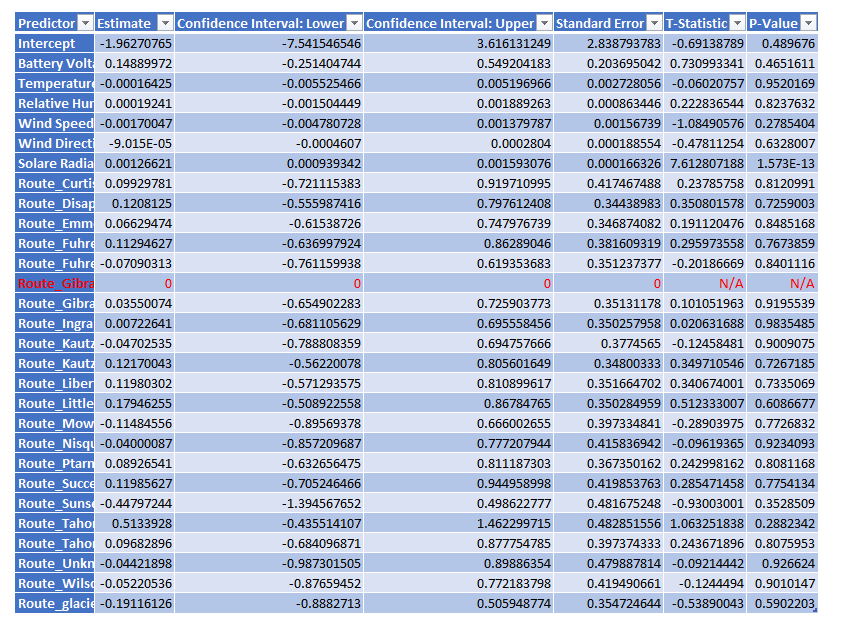


**Data Mining**

For converting the output variable success ratio to a categorical variable, we use a cutoff of 0.4. Any ratio above 0.4 is assumed to be Yes. Because the number of useful climbing statistic rows is reduced to 481, It would be ideal to partition the dataset into 60% Training set and 40% Validation set. The model can be tested using more recent weather data and climbing statistics for measure of performance.

**Linear Regression**

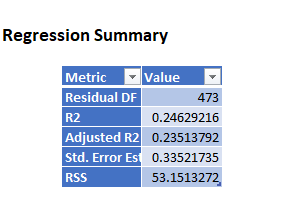
Analysis of Coefficients

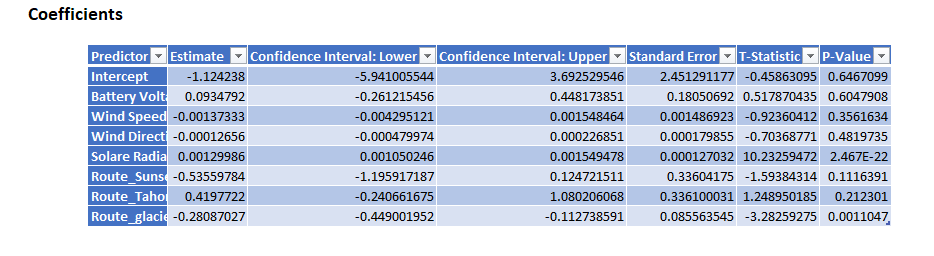


The categorical variable Route is split into binary dummy variable, we also ignore route\_Gibralter because its information would be redundant. We only choose to include variables with low p-values. We do not use the temperature because it is highly related to Average solar radiation and has high p-value. Similarly, Relative humidity is ignored as well. In the p-values for the routes, only routes Sunset Ringraham, Tahoma Cleaver, and glacier only are included because they have statistical significance.

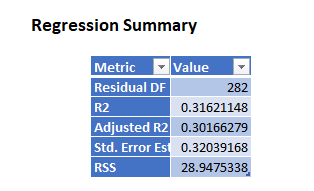
Regression Model

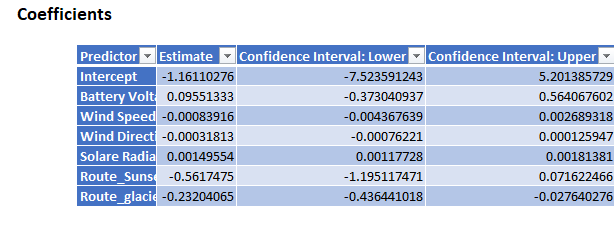
Single Dataset





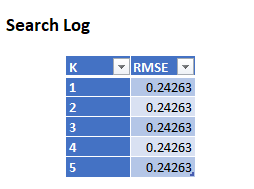
Training and Validation Dataset

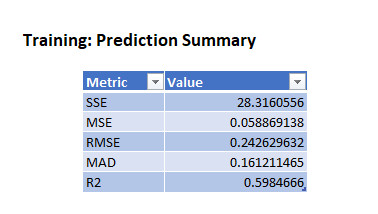




**K-Nearest Neighbors regression**

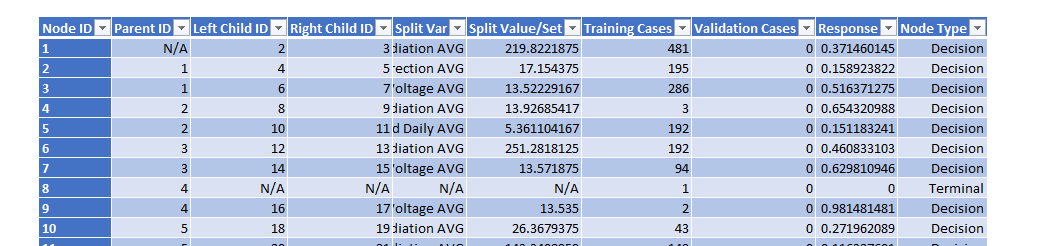
It would be ideal to use K-NN with the entire data as one set because we have limited number of rows. Using the variables from the predictor screening done previously, we get the following model

We use k=3



**Regression Tree**

We fit a regression tree using the entire dataset and fully grow it to try a capture the relationship between the predictors and the response variables. The first few nodes of the fully-grown tree are shown in table form below



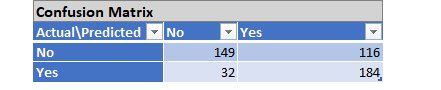
**Classification Data mining**

Our output variable now changes to a categorical one. The numeric output variable is transformed using a cutoff of 0.4 for the success ratio. Any observations above 0.4 will be classified as ‘Yes’ otherwise ‘No’

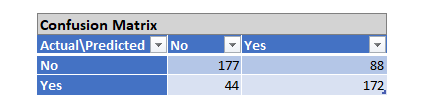
**Discriminant Analysis**

We fit a discriminant analysis model to try and model the data. We choose two models with cutoff probabilities of 0.4 and 0.5 with the success class as Yes. We have to use the entire data as one Training set to ensure accuracy of prediction.

0.4 cutoff confusion matrix



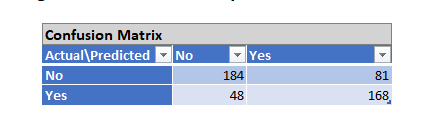
0.5 cutoff confusion matrix



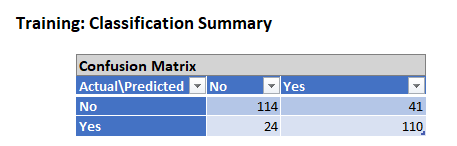
Logistic Regression Model

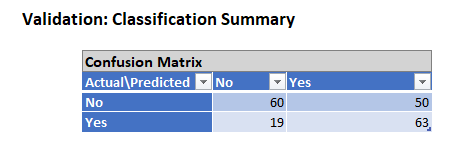
We fit a logistic regression model for classification purposes. We use a cutoff of 0.5 for better performance and we perform analysis on both a single data set, and a data set partitioned into Training and Validation sets.

Single data set confusion matrix

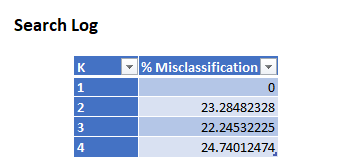


Training and Validation Sets



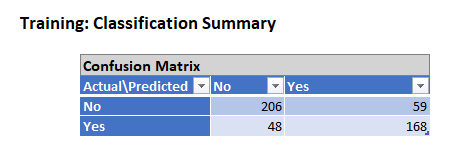


**K-NN Classification**

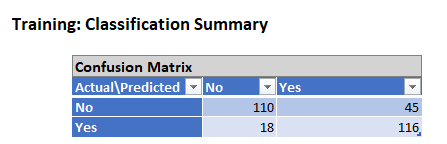


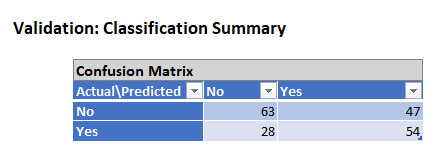
We use k= 3 and cutoff probability 0.5

Single dataset



Partitioned dataset

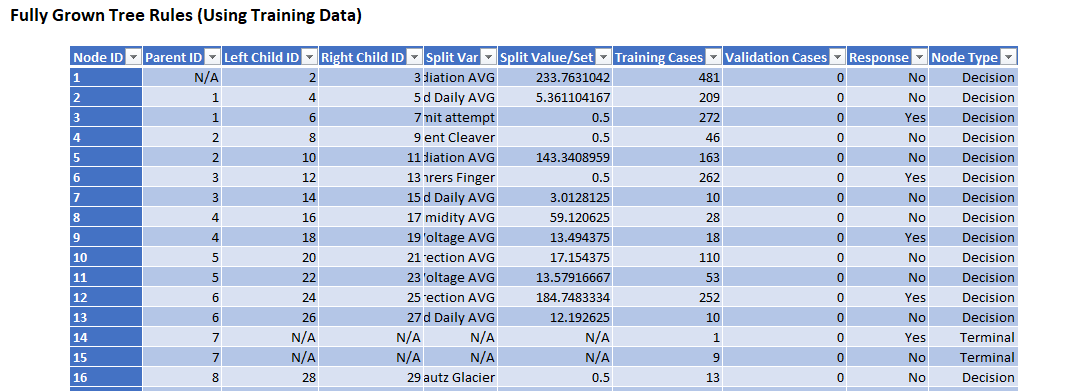


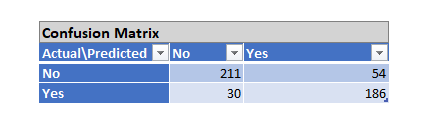


**Classification Tree**

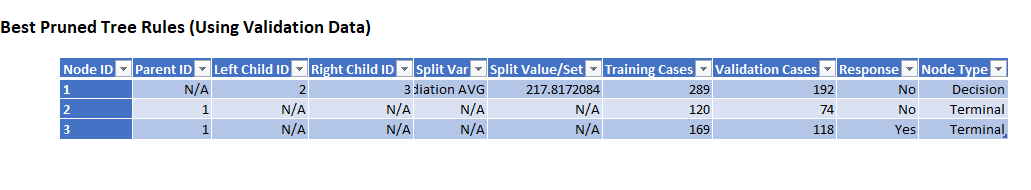
We fit a classification tree to the single dataset and the partitioned dataset

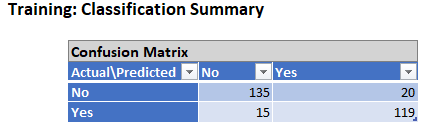
Single dataset





Partitioned dataset

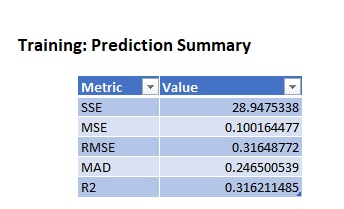


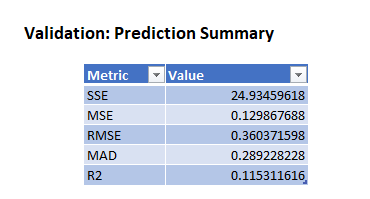


**Performance Analysis**

**Linear Regression**

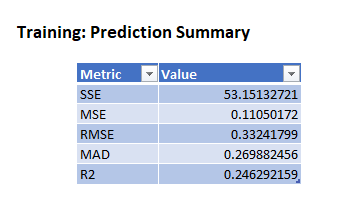
Linear regression with Training and Validation Set





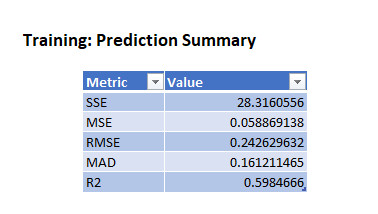
The Linear regression has a bad fit for our model. The R2 value is very low about .316 for the Training set and .115 for the Validation set. Moreover, the scale of our output variable is [0.1], an RMSE value of .316 is high considering this and shows that the model does not have good predictive performance

Linear regression with one dataset



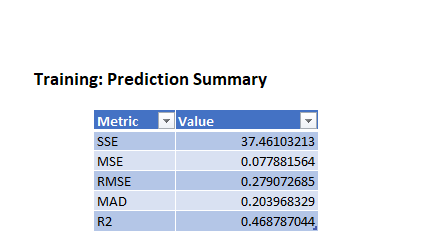
The R2 value for the single dataset is also very low, about .246 and high RMSE of .3324. This suggests the model is not a good fit for prediction through Linear regression.

**K-NN Regression**



The K-NN regression has a comparatively high R2 value. The drawback of this model is the dataset must be considered as one and cannot be partitioned. An RMS error of .242 is still high and needs to be decreased before a K-NN model can be used for prediction. This model has potential use if more data can be provided and included in the analysis

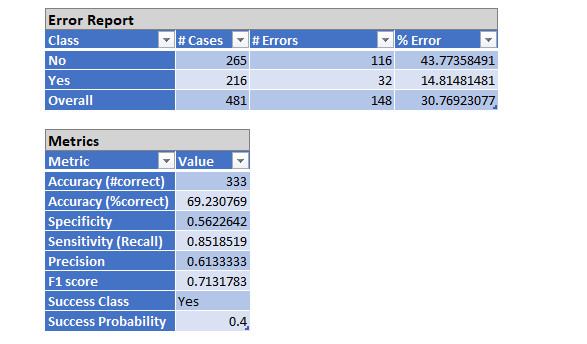
**Regression Tree**



The Regression tree for prediction has an okay fit with R2 at .458. The RMS error of .279 is still high and more data is required before an accurate regression tree can be fitted.

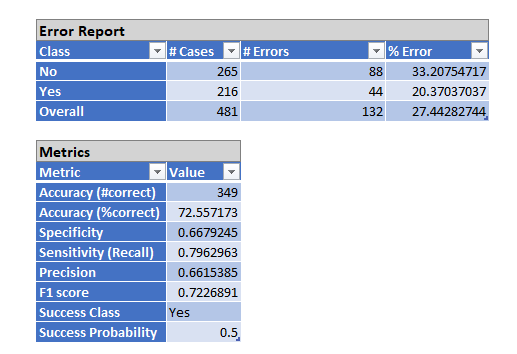
**Discriminant Analysis**

0.4 Cutoff



The Total error for 0.4 cutoff is comparatively high at 30.76%. The 0.5 cutoff has better accuracy and has better applicability.

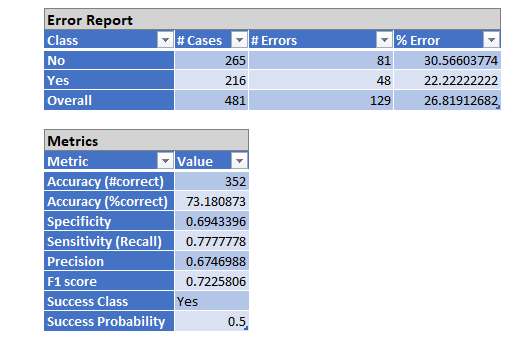
0.5 Cutoff



An error of 27% shows that the Discriminant analysis model has potential for possible application given more data can be incorporated into the model. High sensitivity of 0.79 tells us that the model does a good job in predicting a ‘YES’. Specificity of .66 is slightly low and can be improved upon

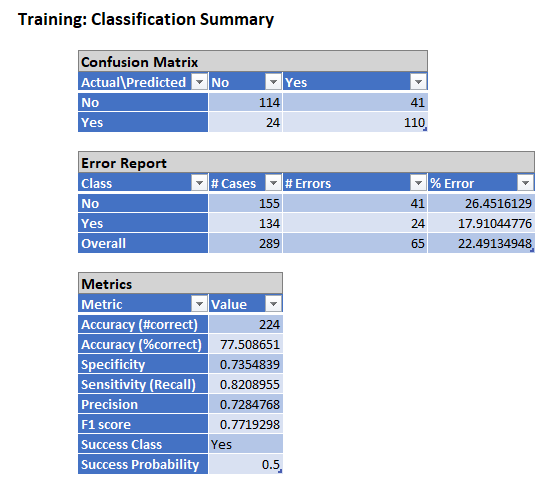
**Logistic Regression**

Single dataset

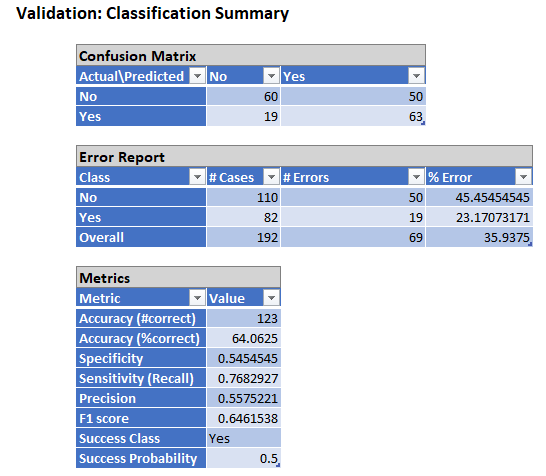


The logistic regression classification model improves on the DA model and reduces the overall error to 26.81%. High sensitivity of 0.77 tells us that the model does a reasonable job in predicted a ‘YES’. Specificity of .69 is acceptable but can be improved upon

Training and Validation Dataset



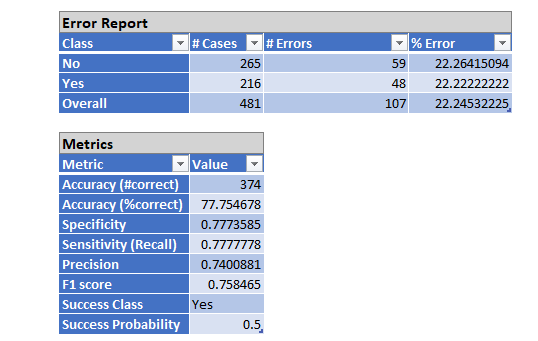
The Training set has good accuracy at 77.5%. A sensitivity of .82 is good and specificity of .73 is acceptable



The Validation set has considerably high error at 35.93% and low specificity of .54 indicates that the model might be overfit to the Training data.

**K-NN Classification**

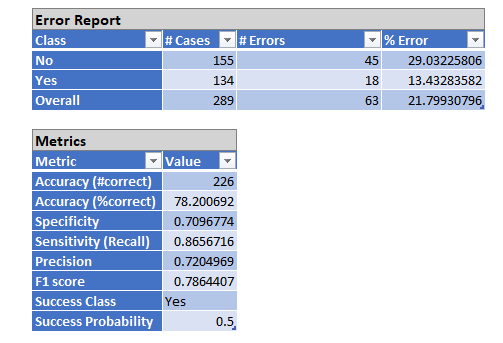
Single dataset



The K-NN classification has good performance with 77.75% accuracy. Along with a high accuracy value, high sensitivity of .77 and specificity of .77 indicates good classification performance which can be improved by adding additional data

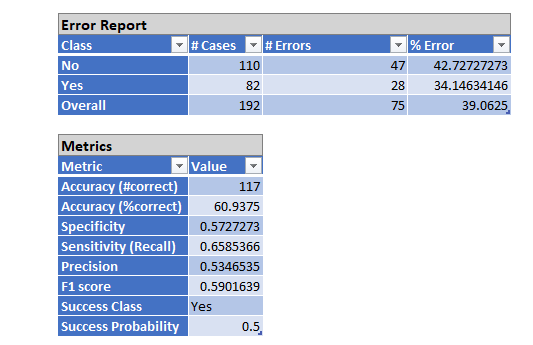
Partitioned dataset

Training



The overall error of 21.79% is good and the model has good sensitivity at .86 but lower specificity at .70

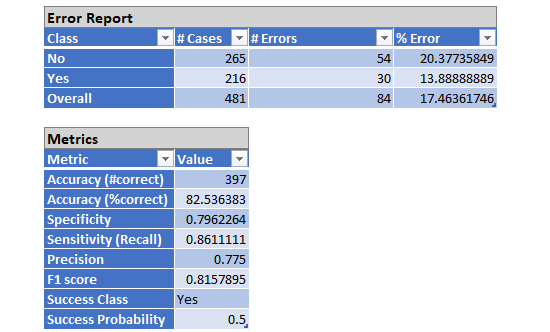
Validation



The high error rate for the Validation set indicates the model might be overfitting to the Training data. But because this is the K-NN algorithm, it could also mean there are not enough records to achieve good classification performance.

Classification Tree

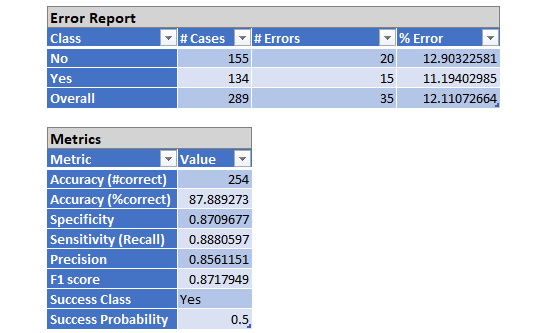
Single dataset



The classification tree has good performance with a comparatively low error rate of 17.46%. Sensitivity of .86 indicates the model does a good job predicting positive outcomes while specificity of .79 is acceptable.

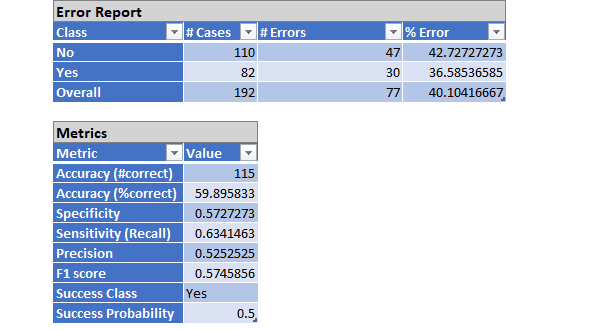
Partitioned dataset

Training



The error for the Training dataset is the lowest among the models at 12.11%. High sensitivity of .88 and high specificity of .87 indicate good classification performance.

Validation



The high Validation error shows the model might be overfitting the data or there are not enough records in the Validation set to achieve good classification performance.

**Project Results**

* The classification tree has the best performance for classification purposes. It is recommended that this model is used for classifying future climbs. The model also shows significant potential for high levels of accuracy if more data was obtained.
* The K-NN regression model has the best predictive performance. With more data the model will do a really good job at predicting the chances for a successful climb.

**Project Impact**

* Hardly any climbs are attempted in the winter. It is advisable to avoid climbing the mountain in the winters regardless of the route.
* The Disappointment Cleaver route is the most commonly used route and has the most successes. However, the route is only frequented in the Fall months.
* The Emmons-Winthrop route is used all year long and has about 50% success rate
* The safest time to attempt a climb is when then weather conditions are good, i.e. good AVR solar radiation and low Wind speeds.
* The safest routes are Emmons-Winthrop, Kautz glacier and Disappointment Cleaver. The Kautz glacier route has a success rate of .55
* The routes that should be avoided are Fuhrer’s Fingers, Gibraltar ledges and Ingraham Direct.
* The best season to make a climb would be Summer leading into Fall, staring in May and running until mid-September. July and August have the highest success ratio of about 0.6. But this also a popular time for climbers with a lot of climbs being attempted.
* The best measure of the weather conditions is the Average solar radiation followed by the Wind speed. The higher the solar radiation the higher the chance of a successful climb. With lower wind speeds, chance of a climb being successful are higher.
* With more data, a robust model using K-NN for prediction and Classification trees for classification can be formulated which can be highly accurate. We already achieve about 80% precision for classification using limited number of records.